Simputation - Grouping fails

January 9, 2019

```
In [1]: library(dplyr, quietly = TRUE, warn.conflicts = FALSE)
        library(simputation, quietly = TRUE, warn.conflicts = FALSE)
        library(naniar, quietly = TRUE, warn.conflicts = FALSE)
```

```
In [2]: data(iris)
        irisNA <- iris
        irisNA[1:4, "Sepal.Length"] <- irisNA[3:7, "Sepal.Width"] <- NA
        head(irisNA)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
NA	3.5	1.4	0.2	setosa
NA	3.0	1.4	0.2	setosa
NA	NA	1.3	0.2	setosa
NA	NA	1.5	0.2	setosa
5.0	NA	1.4	0.2	setosa
5.4	NA	1.7	0.4	setosa

```
In [3]: irisNA %>%
          dplyr::select(Sepal.Length, Sepal.Width, Species) %>%
          group_by(Species) %>%
          summarise(sepalLengthMedian = median(Sepal.Length, na.rm = T),
                    sepalWidthMedian = median(Sepal.Width, na.rm = T))
```

Species	sepalLengthMedian	sepalWidthMedian
setosa	5.0	3.4
versicolor	5.9	2.8
virginica	6.5	3.0

The imputed data below is consistent with what is expected. The NA values are imputed with median values of Sepal.length and Sepal.Width for “setosa” species.

```
In [4]: head(simputation::impute_median(irisNA, . ~ Species))
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.0	3.5	1.4	0.2	setosa
5.0	3.0	1.4	0.2	setosa
5.0	3.4	1.3	0.2	setosa
5.0	3.4	1.5	0.2	setosa
5.0	3.4	1.4	0.2	setosa
5.4	3.4	1.7	0.4	setosa

Create a new iris dataset with randomly shuffled rows. Induce NAs as shown in output below in rows 1 to 3 for setosa and versicolor species.

```
In [5]: set.seed(1)
        iris2 <- iris[sample(150), ]
        iris2[1,1] <- iris2[1:3, 2] <- NA
        head(iris2)
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
40	NA	NA	1.5	0.2	setosa
56	5.7	NA	4.5	1.3	versicolor
85	5.4	NA	4.5	1.5	versicolor
134	6.3	2.8	5.1	1.5	virginica
30	4.7	3.2	1.6	0.2	setosa
131	7.4	2.8	6.1	1.9	virginica

```
In [6]: iris2 %>%
          dplyr::select(Sepal.Length, Sepal.Width, Species) %>%
          group_by(Species) %>%
          summarise(sepalLengthMedian = median(Sepal.Length, na.rm = T),
                    sepalWidthMedian = median(Sepal.Width, na.rm = T))
```

Species	sepalLengthMedian	sepalWidthMedian
setosa	5.0	3.4
versicolor	5.9	2.8
virginica	6.5	3.0

Validating the output below, it is observed that grouping on Species fails in this case and missing values are imputed with median values of "setosa" species. Expected imputed value for Sepal.Length is 5.9 whereas 5.0 is imputed in row number 40, which is first row in the iris2 dataset.

```
In [7]: head(simputation::impute_median(iris2, . ~ Species))
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
40	5.0	3.4	1.5	0.2	setosa
56	5.7	3.4	4.5	1.3	versicolor
85	5.4	3.4	4.5	1.5	versicolor
134	6.3	2.8	5.1	1.5	virginica
30	4.7	3.2	1.6	0.2	setosa
131	7.4	2.8	6.1	1.9	virginica